EXTRACTIVE TEXT SUMMARIZATION USING FUZZY-C-MEANS CLUSTERING

A thesis submitted to the Department of CSE
in partial fulfillment of the requirements for the degree of
B.Sc. Engineering in CSE

Submitted By-
Shafiul Gani
Student ID: 14101213

Supervised By-
Dr. Jia Uddin
Associate Professor
Department of Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University, Dhaka - 1212, Bangladesh

December 2018

DECLARATION

It is hereby declared that this thesis is based on results have found from my research. Materials of work from researchers conducted by others are mentioned in references.

Signature of the Author

Shafiul Gani
Student ID-14101213

Signature of the Supervisor

Dr. Jia Uddin
Associate Professor
Department of Computer Science and Engineering
BRAC University

ACKNOWLEDGMENTS

At first, I want to offer most profound thanks to Almighty Allah for enabling me to effectively entire research work. I would like to express my deepest gratitude to our supervisor Dr. Jia Uddin for his guidance on this project showing me the path of conducting successful research and above all for always being there as mentor. He shared his wisdom with us in analyzing subject matters and at the same time valued our thinking approach to synthesize those topics. His suggestions drove towards better ways of thinking, his reviews enriched me in solving problems. I will forever cherish the memories of working with him. Furthermore, I would like to thank my parents and friends for the support, suggestions, moral help and aids. Lastly, I want to thank BRAC University for the scope and all the support, facilities for doing this research.

TABLE OF CONTENTS

List of Tables

Table **Page**

LIST OF FIGURES

ABSTRACT

Keeping track of the precise information from a large volume of text is an arduous task for human. Test summarization process has become one of the significant research areas for years owing to cope up with the astounding increase of virtual textual material. Text summarization is the process to keep the relevant important information of the original text in a shorter version with the main ideas of the original text for understanding innumerable volumes of information easily within a short period of time. There are two main classifications of text summarization process, Extractive and Abstractive text summarization. Extractive summarization processes by using most important fragments of exiting words, phrases or sentences from the original document. It largely depends on sentence-extraction techniques or sentence-based model. A sentence based model using Fuzzy C-Means clustering has been proposed this research. Six key features including a new feature have been added for the sentence scoring. Performance of the proposed FCM model is evaluated by ROUGE, which has been gauged with the precision, recall and f-measure.The result shows that this FCM model interprets extractive text summarization methods with a less summary redundancy and depth of information and also it shows more adhering and coherent than other previous approaches.

Keywords: Sentence Extraction, Clustering, Summarization.

CHAPTER 1

INTRODUCTION

Human have the inherent ability to comprehend the implicit meaning of a text and summarize the precise information using own words from the most important aspects of the text. That is, they are capable of extracting the salient features to interpret the text in a more concise way. Every day we have to go through a large volume of textual data. For human, it would take years to read, understand and then make summary of this huge amount of text data generated in one day. So it is necessary to have an automatic text summarization technique that would offer the advantage of making precise summary out of the huge amount of text data generated every minute in the virtual world. But it has been quite difficult to make the best use of this progressively increasing amount of data. Search engines have been extracting snippets by using information retrieval systems for making our life easier but those snippets are becoming a larger size documents themselves [1]. Text summarization has been playing a vital role to solve this issue.

Aim of the text Summarization is to prune and filter a large amount of data into a shorter version keeping the most relevant and significant ideas of the original document [2] [3]. There are many advantages of text summarization. A compact summary of a text allows a user to quickly have an overall idea about the text, indexing effectively and also helps to select relevant documents according to ones necessity [2]. Text summarization has been classified into several categories depending on various aspects [1] [2]. Based on input type, it can be categorized in Single Doc-

ument and Multi Document text summarization. Single Document text summarization deals with a single document as input whereas Multiple Document processes an arbitrary length of text document as input. Depending on purpose, it is divided into Generic, Domain-Specific and Query Based summarization. In Generic summarization, no assumption is made about the model and all the contents of the text are treated as homogenous inputs. In Domain-Specific summarization, knowledge of a specific domain is used by the model to make more accurate summary of the text. Query based summarization intends to answer the natural language questions about the input text. Abstractive and Extractive are the two approaches of text summarization based on output type [3]. Abstractive text summarization is referred by many scientists as a human generated summary [4]. Abstractive approaches are much harder.

On the contrary, extractive approaches based on the extraction of sentences with the help of some renounced methods like sentence-based model, word-based model or graph based model. Sentence ranking method is based on some key features and sentence-based model iterates though all the sentences in the document to find out the main ideas. Previous text summarization methods used binary parameters, that is, the value would be either 0 or 1. But this approach does not work properly for most of the situations. In this paper, we try to solve the problem and overcome this scenario by giving the attributes fuzzy quality. We use Fuzzy C-Means clustering algorithm to summarize the text. Fuzzy C-Means clustering algorithm allows data to be member of more than one cluster. They have a certain degree of membership each cluster between 0 and 1 [5]. The closer the value of a data to 1, the more the data is inclined to that cluster and vice versa [5]. The sentences are ranked by the model, based on the statistical characteristics of Information Retrieval methods. The summary is organized as the sentences of the original texts are incorporated in the summary according to importance. A set of analyzers use the extracted information to evaluate the rank of each sentence.Many important features have been introduced in the previous works [3].

Fuzzy C-Means clustering algorithm works in several steps to rank the sentences. First, it creates a Partition Matrix, which holds the value of belongingness of a data to different clusters. Then it iterates using an Objective function to find a Cluster Center. After every iteration, the belongingness value is updated until any stopping criterion is met. This is how the Fuzzy C-Means algorithm works to find the clusters [1] [6]. Here, six of the best features among those included a new future "Sentence Sighlighter Score" has been introduced which can help to improve the performance of sentence ranking with Fuzzy C-Means clustering algorithm [5] [3] [6].

1.1 Motivation

Keeping track of the precise information from a large volume of text is an arduous task for human. Text summarization process has become one of the most significant research areas for years owing to cope up with the astounding increase of virtual textual material. NLP has played a vital role for making our life easier from decades ago [2]. Now, machine learning process has been applied almost in every sector of our life. Daily uses of vast amount of virtual textual data has increased mostly. Every day we have to go through a huge amount of data both for educational, professional and day to day needs. It is a quite difficult job to keep track of the vast amount of data. Apart from this, it requires a huge time for managing the whole document and get the main ideas. Text summarization process has made our life easier. By making a short main ideas we can have the proper idea in a short period of time for a large textual data. Moreover, the application of machine learning has enabled more options for managing the large amount of document data easily. Furthermore, Search engines have been extracting snippets by using information retrieval systems for making our life easier but those snippets are becoming a larger size documents themselves [1]. So, the vast uses of virtual document has increased in a significant way. Science has been always pave our way to make life easier. Though there are many works previously regarding text summa-

rization [5]. But the proposed model has been showing a promising result than before by using Fuzzy C-Means Clustering. Fuzzy C-Means is an unsupervised Machine Learning algorithm in which data can belong to more than one cluster with a certain degree of belongingness is adopted by sentence ranking procedure for sentence extraction purpose.

1.2 Contribution Summary

Previously a number of researchers have done many models for making text summarizer [1] [2]. Still, we have to face many difficulties to get the exact outputs we cherish. There are few supervised, graph based model have been applied for making text summarizer. Yet the best result is found from unsupervised methods. The basic task to make a text summarizer to find out the main keywords which holds the main idea. So, the main and first most challenging task was finding the best features. Though there are many features for scoring have been introduced but still the output is not satisfactory yet. Comparing all the existing features, five best features have chosen with a new feature called Sentence Highligher Score for our model. includes the connection among sentences correlation is imperative for the outline as the sentence frequently alludes to the past or the following sentence. On the off chance that we consider just the connection of a sentence with the past sentence at that point sentences beginning with connectives, for example, this, those, moreover, however, such, although etc. related with significant data reserved sentences. Apart from this, highlighted bullet points, quotation, bold words portrays a significant meaning of that documents. So, the most highlighter found in a sentence has a greater chance to represent an important idea of the document. Therefore, this feature includes the value of a sentence by adding the frequency of highlighter in a sentence compared to the maximum gain highlighter score of a sentence from the document. Moreover, choosing the algorithm for extracting best summary was also challenging. There are various existing models available. We have found the result from unsupervised method has significantly remarkable. We have used the algo-

rithm Fuzzy-C-Means for extracting best sentences. The reason behind choosing Fuzzy-C-Means, it is an unsupervised soft computing technique which creates clusters based one data importance. Above all it was always a big challenge for a machine to understand the value apart from 0 or 1. In Fuzzy-C Means a data can be stored based on the scoring value not only just 0 or 1 but any values from 0-1 [5]. Above all, in the regular methods a data can be clustered in only cluster which scored between 0 or 1 .But, here data can belongs to one or more clusters according to their scoring value. So, the proposed model provides a clear promising summary altogether with the features than many other models.

1.3 Thesis Outline

The remain of the report is outlined as follows. Chapter 2 portrays the previous works done in the field of text summarization techniques. Furthermore, chapter 3 describes the proposed FCM model. Then, Chapter 4 represents illustration of the experimental results based on a versatile dataset and the performance comparison of other renowned models. Finally, chapter 5 concludes the paper mentioning about the research experiences and future works.

CHAPTER 2

LITERATURE REVIEW

Text summarization has been a research are for more than 70 years. There are many proposed models have been found in this area. Everyday the process has been improving. Still, there is a need of more human like summzarizer for making the best use.

2.1 A Brief History of Text Summarization Technique

The very first text summarization technique was introduced by Luhn [4] back in 1958 using the the thematic feature, Term- Frequency. This paper depicts some exploratory research on programmed strategies for getting abstracts. The framework laid out here starts with the archive in machine-coherent frame and continues by methods for a modified testing process equivalent to the checking a human purser would do. The key sentences are then identified to fill in as pieces of information for making a decision about the character of the article. It has been almost 70 years, text summarization technique has been a research concern and researchers have introduced many genres to represent the best outcome of text documents [7].

Again in 1958, sentence location for assessing sentence importance was introduced by Baxndale [8]. The examination was exploratory in nature and was roused by an undertaking to decrease the lopsided work required to process the topic of distributed writing. This abstracting and ordering of info records speaks to somewhere around 80 percent of the exertion of momentum writing looking frameworks as against 20 percent given to re-

covery. The need to address the lopsidedness is self-evident. The work was done on the suspicion that systems for perusing straightforwardly from a record will eventually be formulated. Three techniques were contrived for checking a record to remove the fundamental substance of printed matter. One was a filtering of point sentences; the second was a grammatical erasing process; and the third was a programmed choice of expressions. In each example the subsequent vocabulary was positioned by recurrence circulation. The lists removed by every strategy, and also the recurrence designs coming about, were then looked at.

Practical evidences mentioning difficulties innate in the perfect summary concept was visible in the work of Rath et al. [9] in 1961. The examination demonstrates that both the human chose sentences and the program chosen sentences contrast essentially from irregularity. There is an extensive variety of individual contrasts between the human subjects, while the 5 strategies yield little contrasts in the sentences picked. There was next to no understanding between the subjects and the machine techniques on the sentences chosen as being agent. Auto-abstracting procedures dependent on high-recurrence words demonstrate to a great degree little variety among themselves in the determination of sentences. Human determination of sentences, albeit less factor than chance anticipation, is impressively more factor than machine techniques. There was next to no assertion between the subjects and machine techniques in their choice of delegate sentences.

On the other hand syntactic analysis [10] in text summarization introduced the notion of entity-level approaches. Furthermore, the use of Bayesian classifier [11] introduced a probabilistic approach for the sentence selection for summarization. This paper presents a programmed content outline way to deal with beat the challenges in the current rundown approaches. Here, Gullible Bayesian Arrangement approach is used to distinguish the fundamental watchwords from the content. Bayes strategy is machine learning technique to gauge the distinctive catchphrase includes in a content and recovers the watchword from the info dependent

on this data. The highlights are commonly autonomous and circulated. Scoring is evaluated for the recovered sentence to process the word recurrence. The blend of this Credulous Bayesian, scoring and timestamp idea enhances the rundown precision. The proposed outline strategy accomplishes better inclusion and cognizance utilizing the Guileless Bayesian classifier and the idea of timestamp, thus it naturally disposes of the excess in the information archives.

Moreover, from the late 90s are bushy path and aggregate similarity [12] has been used for text summarization. Erkan proposed LexRank [13] algorithm in 2004 inspired by the graph based model. A sentence availability grid was figured dependent on the cosine closeness work. On the off chance that the cosine comparability among the two sentences goes past an explicit edge, at that point the edge was added to the network grid.

2.2 Related Works

Fuzzy sets [14] provide a solution in text summarization technique by denoting a parameter to measure the degree ambiguity in a context. In spite of a number of works done in fuzzy logics based text summarization, the Fuzzy C-Means (FCM) clustering is hardly explored in this area [3] [15] [16].

Graph based models have been one of the core inspirations behind the exploration in extractive [17] summarization approaches. In A four dimension Graph Model for Automatic Text Summarization [17] a graph model has been utilized to form extractive summary.In later works, sentence based extractive summarization methods have been proven to be more efficient and less time and space consuming than graph based and word based models [18].

Zheng et al. [19] used applied relations of sentences for multidocument summarization. This idea was made out of three noteworthy components.

They were idea bunching, sentence idea semantic connection, and synopsis age. The sentence idea semantic connection was achieved plan the sentence idea chart. The diagram weighting calculation was rushed to acquire the positioned weighted ideas and sentences. At that point, bunching was connected to evacuate the excess and rundown age was directed to recover the useful synopsis.

Glavad and Snajder [8] proposed event graphs for data recovery and multidocument synopsis. An event based archive portrayal approach was acquainted with channel and structure the insights about the events clarified in the content. Principle based models and machine learning were incorporated to separate the sentence level events and assess the fleeting relations among them. A data recovery approach was utilized to gauge the similitude among the reports and inquiries by assessing the chart bits crosswise over event graphs.

Ferreira et al. [1] planned a multidocument synopsis display dependent on linguistic and statistic treatment. This methodology separates the significant worry of set of reports to maintain a strategic distance from the issues of this sort of outline. It was acquired with the assistance of grouping calculation which utilizes the measurement similitudes and semantic treatment.Meena and Gopalani [20] proposed a framework that is domain independent for text summarization.

Sankarasubramaniam et al. [21] presented a content synopsis utilizing Wikipedia. This methodology develops a bipartite sentence idea diagram and the input sentences were positioned dependent on the iterative updates. Here, a customized and question centered synopsis was considered for client inquiries and their interests. The Wikipedia based multidocument summarization was proposed, which allows incremental streaming summarization. K. Vimal Kumar, Divakar Yadav, in their paper primarily laid accentuation above all on the Hindi content summarization which depicts different highlights utilized for the Hindi synopsis [22].

However, in this paper, a novel approach is proposed using FCM algo-

rithm for sentence extraction for generating summaries. The FCM algorithm is an unsupervised soft computing technique that uses fuzzy sets and fuzzy partition matrix to denote the membership of an element across multiple clusters [5]. The membership value of a sentence in the partition matrix for determining the significance of the sentence is the idea behind proposed FCM model as well as generating relevant significant summary of the textual document.

2.3 Machine Learning Approaches

Machine Learning is a group of algorithms that empowers programming application to produce likely the most noteworthy precise in computing results without being unequivocally modified [1]. The common proposition of machine learning is to create or develop algorithms that receive input data and make statistical analysis for predicting an output. We have used machine learning techniques for automated text summarization. It is broadly used in applications like self-driving cars, spam filtering, web search, graphic recognition system, identification of document genre, authorship attribution, automated essay grading, classification of news articles etc. [2] [4].The imminent need to access the rising availability of documents in digital form, the content-based document management tasks have earned a remarkable status in the information systems area. It is used broadly in NLP for categorize the text and processes for further uses. Text summarization technique is also found by using machine learning algorithms. There are basically two main classification of machine learning algorithms supervised learning and unsupervised learning. In this research unsupervised learning algorithm Fuzzy-C-Means Clustering algorithm has been used for the text summarizion model. And has been compared to the other popular unsupervised learning algorithm for comparing our model.

2.3.1 Fuzzy Logic

Fuzzy logic is an approach to computing based on degrees of truth rather than the usual true or false (1 or 0) Boolean logic on which the modern computer is based. The idea of fuzzy logic was first advanced by Dr. Lotfi Zadeh of the University of California at Berkeley in the 1960s. Natural language is not easily translated into the absolute terms of 0 and 1.It may help to see fuzzy logic as the way reasoning really works and binary or Boolean logic is simply a special case of it. Fuzzy logic includes 0 and 1 as extreme cases of truth (or the state of matters or fact) but also includes the various states of truth in between so that, for example, the result of a comparison between two things could be not tall or short but of tallness. Fuzzy logic seems closer to the way our brains work. We aggregate data and form a number of partial truths which we aggregate further into higher truths which in turn, when certain thresholds are exceeded, cause certain further results such as motor reaction. Fuzzy logic is essential to the development of human-like capabilities for AI, sometimes referred to as artificial general intelligence.

2.3.2 K-Means Algorithm

K-means clustering is a kind of unsupervised learning algorithm and used for un-leveled datas. The objective of this algorithm is to discover bunches in the information. The variable k is is iteratively added new data points based on the features. The clusters hold the significant data considering similar significance in the document. The centroid of the cluster compares the values and results in grouping of similar significance data in same group.

$$J = \sum_{j=1}^{K} \sum_{i=1}^{N} \| x_i - c_j \|^2 \tag{2.1}$$

Here, k=number of cluster, j=objective function, n=number of cases, $C_j = centroid\, for\, cluster j$.

2.3.3 Minibatch K-Means Algorithm

MiniBatch K-means is a variant of kmeans bunching. Time unpredictability of kmeans clustering is is too high and take a long time to measure and fix the grouping of data. Then again, MiniBatchKMeans calculation takes little clusters (arbitrarily picked) of the dataset for every emphasis. It then assigns a cluster to each data point in the batch, depending on the previous locations of the cluster centroids. It then updates the locations of cluster centroids based on the new points from the batch.

2.3.4 Graph Based Model

Graphs are represents the mathematical similarity and graphical connection between objects [16]. Graph based methods are generally non-parametric. It applied for the fixed dataset and embedded with a graph . A weighted graph, for each data is known as vertex [4].

CHAPTER 3

PROPOSED FUZZY C-MEANS CLUSETERING MODEL

3.1 Block Diagram of Proposed Fuzzy C-Means Clusetering Model

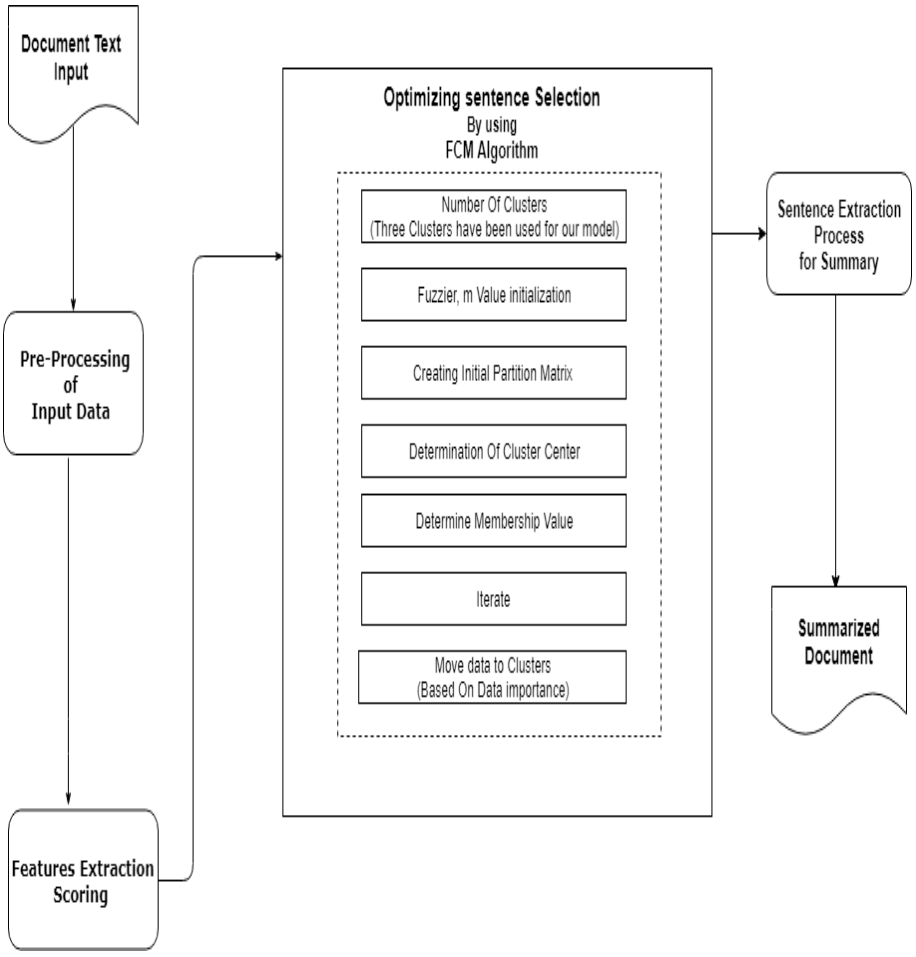

Figure 3.1: Block Diagram of Proposed Fuzzy C-Means Clusetering Model

3.2 Dataset

News article always follows a pattern of writings which helps most to understand the fact easily with proper information. Basically, a regular news article consists of five different parts [23] including headline, byline, lead paragraph, explanation and additional Information. Headline represents the highlights of the article, byline refers about the writer and other parts continue with the elaborate information. For this experiment we have chosen the dataset of CNN news articles as it has been used previously for various purposes [6]. The reasons behind choosing CNN articles included mostly the neat and clean writing following a proper structure of news article. Apart from those cleanliness the heighted part of the articles helps better to understand the main idea easily. Moreover, the bullet highlighted parts of the articles represent mostly the main ideas. Fifty articles of the CNN news have been chosen for our experiment purpose. The articles contains the idea properly above all in a format which helps to identify the main ideas from the article properly and the result shows a clear idea about the effectiveness of the selected article.

3.3 Pre-Processing Task

The pre-processing phase of input data through NLTK a python library included tasks of:

Splitting: The splitting process of data includes two parts included Paragraph splitting and Sentence splitting. The task of paragraph splitting is conducted by finding identifier between two paragraphs mostly using a regular expression. The next part consists of splitting the text into paragraph using a built in l NLTK sentence tokenizer.

Stop Word Removal: The words which hardly impacts the article are regarded as stop words. The task Stop Word Removal includes using stopwords from nltk.corpus. This removal helps to a more clean computational data for extraction purpose.

Parts of Speech Tagging: This process includes the tagging of parts of speech of the each tokens with various tag like noun, pronoun, verb, adjective etc .

Stemming and Lemmatization: Stemming is the way toward decreasing enunciation in words to their root structures, for example, mapping a gathering of words to a similar stem regardless of whether the stem itself is certainly not a legitimate word in the Language. The procedure of lemmatization alludes to the mapping of action word structures. For example, the vast tense and things into the particular shape so the type of the word can be known.

3.4 Feature Extraction

3.4.1 TF-IDF Score

This is the specific first component presented by Luhn [4] in 1958 for sentence extraction and measure the uniqueness of a sentence [1] [7]. Tf-Idf is a surely technique to assess how critical is a word in a report. Tf-Idf are is an extremely fascinating approach to change over the printed portrayal of data into a Vector Space Model (VSM)[4] .The number of times a term is found in a report is called its term frequency. Inverse document frequency decreases the heaviness of terms that happen often in the record set and builds the heaviness of terms that happen once in a while. TF-IDF condition is produced on the idea that the words that are just found in a specific archive possibly holds the important data of that specific document. In this analysis each sentence has been considered as a record and each TF-IDF esteem has been standardized with the most extreme acquired [15]. TF-IDF score has been estimated utilizing Equation (3.1) and condition (3.2) for normalizing the estimations of sentences,

$$TF-IDF(term) = frequency(term) \times log\frac{frequency(term)}{No.of\,Sentences} \qquad (3.1)$$

For a sentence S_i,

$$TF-IDF(Score) = \frac{Sum\ of\ TF-IDF(term)\ in\ S_i}{Max\ sum\ of\ TF-IDF(term)\ in\ a\ Sentence} \quad (3.2)$$

3.4.2 Sentence Length Score(SLS)

It is used for filtering the least and maximum sentence size as they are considered not that important for sentence ranking [1]. The length of the sentence represents the significance of sentence in synopsis. Sentences that are long and short are not appropriate for rundown. The rationale behind this component is that typically the short sentences on a report are the sentences that just might contain only single words, creator names and exclamatory expressions though long sentences are normally found in grabbers and citation which are obscure identified with the essence of the archive. So, sentences that are long will have superfluous data which isn't valuable for synopsis of archive. While, sentences that are too short don't give quite good data for summary.

For a sentence S_i Length Ratio(SLS),

$$SLS(S_i) = \frac{Length\ of\ S_i}{Mean\ Length} \quad (3.3)$$

3.4.3 Numerical Value Score(NVS)

Sentences containing numerical values are viewed as vital as they can possibly contain valuable information [1] [16] [23].The numerical value in a sentence contains a significant data rather than others is observed very often. This influences that sentences that contain numerical information more significant than the others. For a sentence S_i Numerical Value Score(NVS),

$$NVS(S_i) = \frac{No.\ of\ Numerical\ Data\ in\ S_i}{Length\ of\ S_i} \quad (3.4)$$

3.4.4 Tittle Sentence Score(TSS)

Sentences first and foremost characterize the subject of the archive though sentences at last finish up or outline the record. The positional estimation of a sentence is determined by doling out the most elevated score an incentive to the primary sentence and the last sentence of the report [17]. Second most elevated score esteem is doled out to the second sentence from beginning and second last sentence of the report. Remaining sentences are allocated a score estimation of zero.

For a sentence S_i,

$$TSS(S_i) = \frac{Sum(UniqueValScore)}{Max(Sum(Scoreina Sentence))} \tag{3.5}$$

3.4.5 Proper Noun Count Score(PNCS)

A sentence containing high numbers of proper noun is considered more important than other sentences [1] [10]. It refers the main ideas in the document the sentences which contains higher number of proper nouns.

For a sentence S_i,

$$PNS(S_i) = \frac{No.of Nouns in S_i}{Max No.of Proper Nouns in a Sentence} \tag{3.6}$$

3.4.6 Sentence Highlighter Score(SHS)

The novelty of this proposed model is this feature Sentence Highlighter Score which includes the connection among sentences correlation is imperative for the outline as the sentence frequently alludes to the past or the following sentence. On the off chance that we consider just the connection of a sentence with the past sentence at that point sentences beginning with connectives, for example, this, those, moreover, however, such, although etc. related with significant data reserved sentences. Apart from this, highlighted bullet points, quotation, bold words portrays a significant meaning of that documents. Therefore, this feature includes the value of a sentence by adding the frequency of highlighter in a sentence compared to the maximum gain highlighter score of a sentence from the document.

3.5 Fuzzy C-Means Clustering

Fuzzy C-Means clustering algorithm is an soft computing method for the analyzing data and construction of models. Fuzzy sets was presented by Dunn [19] in 1973.Later on, enhanced by Bezdek [20] back in 1981..In 1965 Zadeh [13] used fuzzy sets in the Fuzzy C-Means clustering. FCM provides more than a hard clustering. The main logic behind FCM is one data can belong to the multiple clusters rather than just in a single cluster[2]. Whereas Fuzzy Set follows,

For a set A and an element X,

$$X : X \in A | X \notin A$$

For an indicator function,

$$I_X = 0, | X : X \notin A$$

$$I_X = 0, | X : X \notin A$$

Normal set hypothesis, an announcement takes double esteem, implying that an announcement can be either valid (1) or false (0), in fluffy set hypothesis the announcement can take parts, which implies it very well may be valid with an enrollment estimation of 0.5, 0.55, 0.1, and so forth.So, in FCM clustering the value between 0-1 of an object can be initialized in different clusters which helps more to gather the idea of not only binary clustering only two different mode like 1 or 0.Fuzzy C-Means clustering algorithm works in several steps to rank the sentences. First, it creates a Partition Matrix, which holds the value of belongingness of a data to different clusters. Then it iterates using an Objective function to find a Cluster Center. After every iteration, the belongingness value is updated until any stopping criterion is met. This is how the Fuzzy C-Means algorithm works to find the clusters.

3.5.1 Partition Matrix

The fuzzy C partition of a set S is represented by U, S .Where,

$$Partition\,Matrix, U = \left(\left(\mu_{ij}\right)\right)_{N \times C} \tag{3.7}$$

The partition matrix U must satisfy the following constraints,

- $0 \leq \mu_{ij} \leq 1$
- $\sum_{j=1}^{c} \mu_{ij} = 1,\ for all\ i = 1,2.....N$
- $0 < \sum_{j=1}^{c} \mu_{ij} \leq$ N , for all j=1,2 C

The primary constrain is the scope of the membership value for every component. The second requirement represents that for every one of the components in the set S, the total of their enrollment to the groups must be equivalent to 1.Lastly, the third constrain illustrate that for each bunch, the sum of the membership values for all component of S must be entirely between 0 to N.

3.5.2 Objective Function

The FCM algorithm is focused on until any termination criterion is met, attractively minimize the value objective function J and denoted as,

$$J = \sum_{i=1}^{N} \sum_{j=1}^{C} {\mu_{ij}}^{m} \| x_i - c_j \|^2 \tag{3.8}$$

Where x_i is the data element and c_j is the cluster center.$The fuzzier, m \in [1, \infty)$.

3.5.3 Cluster Center

Cluster center c_j calculation formula is,

$$c_j = \frac{\sum_{i=1}^{N} \mu_{ij} \cdot x_i}{\sum_{i=1}^{N} \mu_{ij}} \tag{3.9}$$

3.5.4 Membership Value

The formula for updating the membership values, i_jf the partition matrix is,

$$U_{ij} = \frac{1}{\sum_{k=1}^{C} (\frac{\|x_i - c_j\|}{\|x_i - c_k\|})} \tag{3.10}$$

3.6 Work flow

3.6.1 Input Data

After the splitting sentences of the document thorough NLTK [18] input data is ready to be clustered and each sentences represented as a 6 ,Äì dimensional vector. Hence, the document of N sentences is represented as $N \times 6$-dimensional matrix, is our input matrix.

3.6.2 Clusters

Clustering is the process of grouping similar kind of data in a particular group or cluster.For the classification of the input document there is 3-clusters used for this experiment based on sentence ranking importance.

- Cluster 01-High Importance.
- Cluster 02-Mid Importance.
- Cluster 03-Low importance.

For the proposed FCM Model the cluster number has been selected as three for categorization of the importance of the sentences of the document for creating summary.The higher value score of a sentence remarks the position in cluster one to others in descending order.

3.6.3 Fuzzier

In Fuzzy C-Means clustering a fuzzier value has to assign for generating the membership value as well as for the comparison of the value of objective function from the centroid .

The fuzzier value ,$m \in [1, \infty)$.

Changing the fuzzier value m seven experiments were experimented and the values were 1, 2, 2.5, 3, 3.5, 4, 5.

3.6.4 Initial Partition Matrix

Formula of initial partition matrix calculated,

$$U^{(0)} = \left(\left(\mu_{ij}\right)\right)_{N\times C} \tag{3.11}$$

3.6.5 Termination Criterion

Two terminations criterion,

- Error limit, e=.0001
- Maximum Iteration =1000

3.6.6 Iteration

1. Cluster centers denoting using equation (3.9)
2. Objective function denoting using equation (3.8)
3. Update partition matrix using equation (3.10)
4. Check termination Criterion and stop.
5. Else back to step 1.

3.6.7 Sentence Extraction for Summary

The General idea of choosing sentences from their membership value to the higher cluster (j=1). As FCM Algorithm breaks down the full membership of a sentence to single cluster to all the three clusters. The algorithm for selection sentence $S_{i=1,2...N}$, based on their membership in

$$U^t = \left(\left(\mu_{ij}\right)\right)_{N\times C} \tag{3.12}$$

The clusters of the experiment eventually can be converted to hard clustering.For generating summary, based one the highest highest cluster membership value of the sentence and with respect to the features link the sentences have been classified in clusters.The highest scorer sentence of a cluster clustered is regarded as most important and comparing to the centroid the most relevant sentences have been extracted.And for customizing summary length there is a input and corresponding value of the length is to be considered while generating summary.

CHAPTER 4

EXPERIMENTAL ANALYSIS

4.1 ROUGE Analysis

Evaluating summary is quite difficult task .A summary may have different types of summaries with various sentences but the idea is illustrated perfectly in all cases. So, one perfect summary of a document cannot be said. ROUGE (Recall Oriented Understudy for Gisting Evaluation) has become a popular platform and standard method of evaluating summaries [24]. As it compares n-gram statistics approach to measure the precision, recall and f-measure of a summarizer.

The equations for measuring,

$$recall, r = \frac{(Human\,Generated\,Summary) \cap (Generated\,Summary)}{(Generated\,Summary)} \tag{4.1}$$

$$precision, p = \frac{(Human\,Generated\,Summary) \cap (Generated\,Summary)}{(Human\,Generated\,Summary)} \tag{4.2}$$

$$f - measure, f = \frac{(2 \times r \times p)}{(r + p)} \tag{4.3}$$

4.1.1 Feature Based Comparison of the Model

The feature based comparison of the model has been experiments by separating the features and the result shows a promising output for the six features together.

- TIF = TF-IDF Score, SLS = Sentence Length Score
- NVS = Numerical Value Score, TSS = Tittle Sentence Score
- PNC = Proper Noun Count Score, SHS = Sentence Highlighter Score

The gauges,
r=recall, p=precision, fm= f-measure

Using Different Features Based Result

Table 4.1: Result From Using Features TIF + SLS + NVS +TSS

Value	r	p	fm
Max	0.74	0.5	0.6
Avg	0.47	0.30	0.37
Min	0.14	0.06	0.08

Table 4.2: Result From Using Features TIF + SLS + NVS +TSS + PNC

Value	r	p	fm
Max	0.75	0.6	0.67
Avg	0.52	0.43	0.47
Min	0.29	0.15	0.19

Table 4.3: Result from using all features together

Value	r	p	fm
Max	0.77	0.6	0..69
Avg	0.55	0.48	0.52
Min	0.21	0.19	0.21

4.1.2 Different Fuzzier Value Based Comparison

Changing the fuzzier value m seven experiments were experimented and the values were 1, 2, 2.5, 3, 3.5, 4, 5. The results shows different progress rate the best progress can be noticeable in the value of m=2.5.

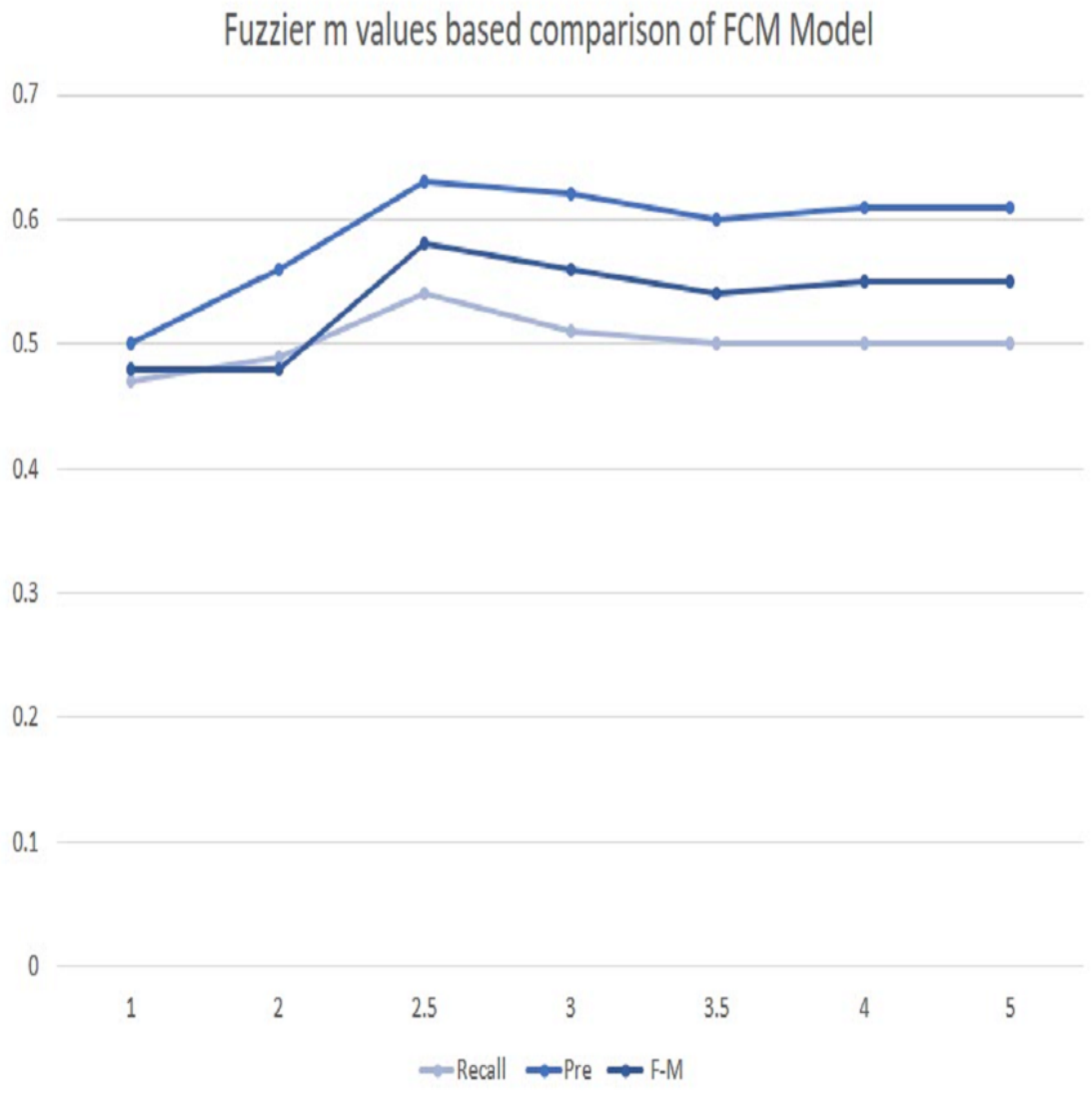

Figure 4.1: Different Fuzzier Value Based Comparison

4.1.3 Similar Approaches Based Comparison on CNN Dataset

K-means is popular unsupervised algorithm which creates clusters for un-leveled dataset and iterate to make group of similar data. But the time complexity of K-Means is too high. Moreover, the f-measure shows a lower result then our model. Minibatch K-Means is an upgraded version of K-means algorithm and this approach is faster than K-Means. Lastly, fuzzy logic which is another popular approach for text summarization and it shows a moderate result scoring 0.47. Our model scored 0.53 which is a promising f-measure and determines the accuracy of finding main ideas and compared to the gold class summary.

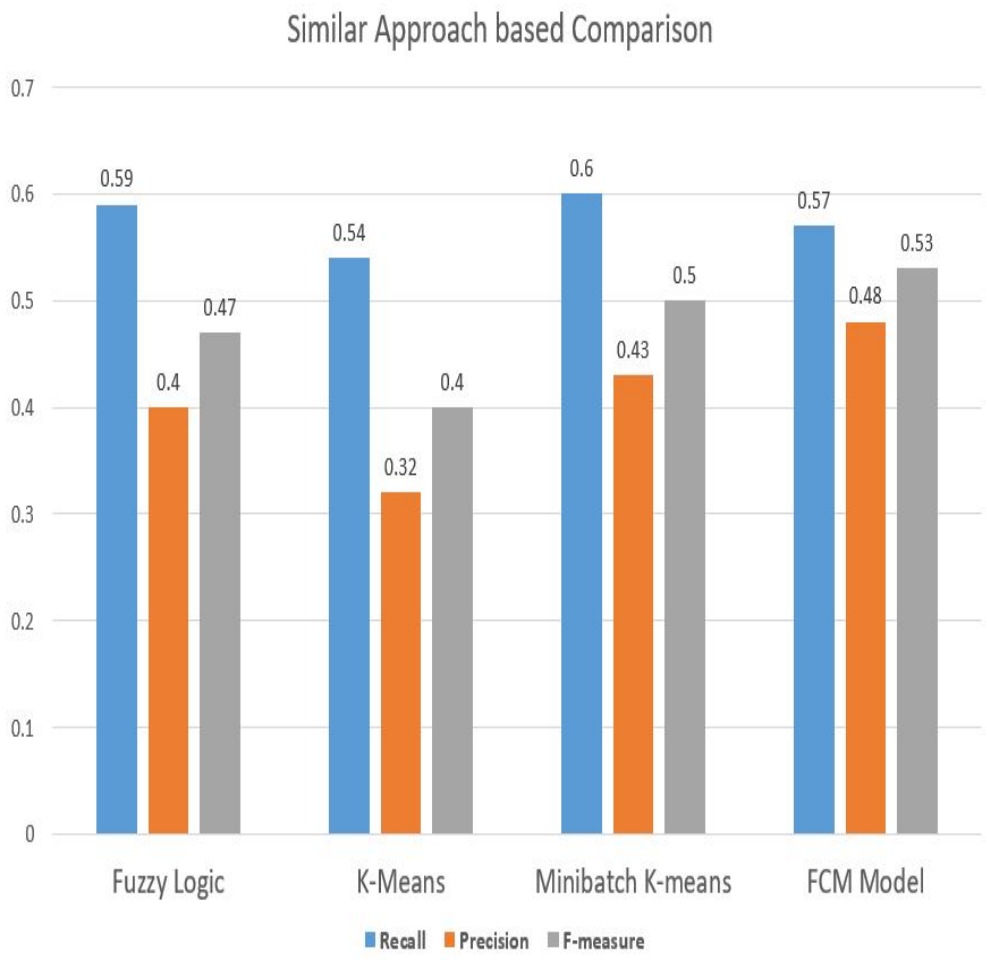

Figure 4.2: Similar Approaches Based Comparison on CNN Dataset

4.1.4 Graph Based Model Comparison

Graphs are represents the mathematical similarity and graphical connection between objects [16]. Graph based methods are generally non-parametric. It applied for the fixed dataset and embedded with a graph . A weighted graph, for each data is known as vertex [4]. It shows a moderate result than our proposed model.

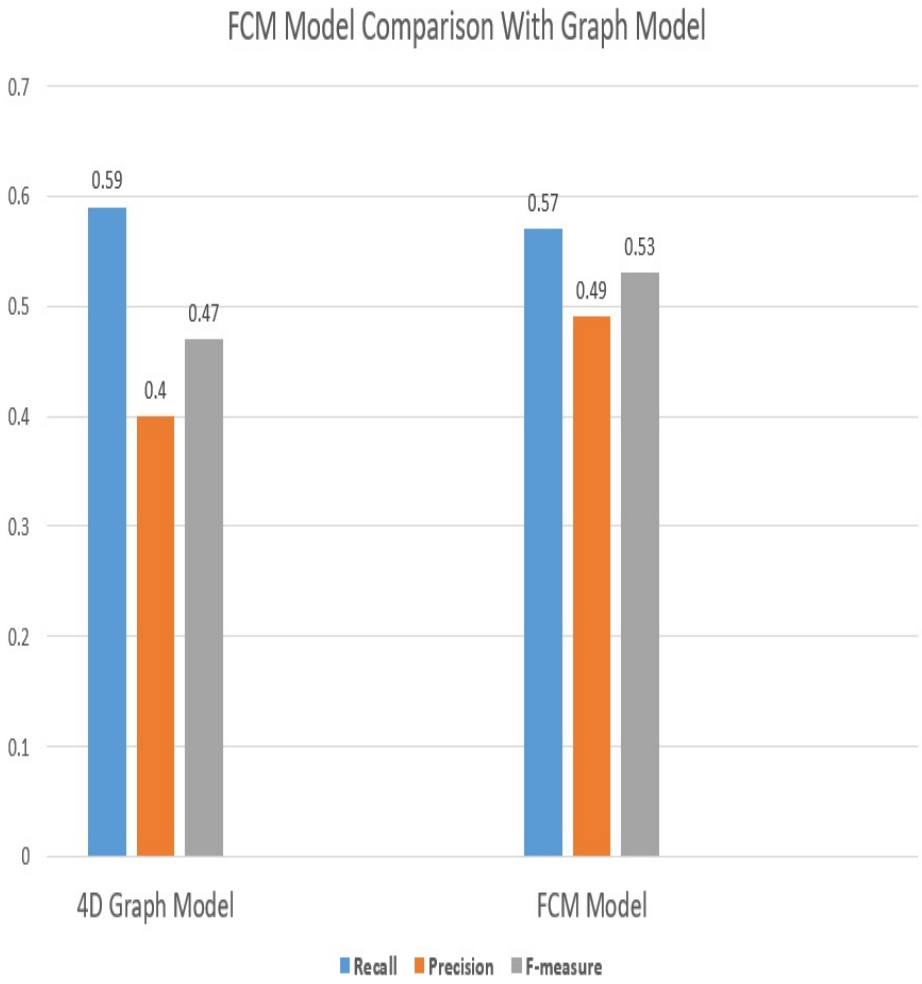

Figure 4.3: Graph Based Model Comparison

4.2 Sample Output

Original Article-You have two ages, chronological and biological. Here's why it mattersReminiscent of a scene from "The Social Network," the whiteboard in researcher and professor Morgan Levine's Yale Medical School office is covered in a series of letters and numbers. She clicks the red cap back onto the dry erase marker and steps back to admire her work.In front of her, the equation stretches across multiple lines, taking up much of the surface. This algorithm represents a new way of thinking about age. "In my lab, we work on a lot of different types of aging measures," Levine said. "One of the most recent ones is based on blood measures you get at your normal doctor's appointment. We basically take those and combine them using different algorithms to get what we call someone's phenotypic age or biological age." Essentially, everyone has two ages: a chronological age, how old the calendar says you are, and a phenotypic or biological age, basically the age at which your body functions as it compares to average fitness or health levels. "People of the same chronological age aren't all at the same risk for developing cardiovascular disease or cancer or even dying," Levine said. "What [the biological age] does is actually give us a better idea of Levine and her team identified nine biomarkers taken in a simple blood test that seemed to be the most influential on lifespan. The biomarkers include blood sugar, kidney and liver measures, and immune and inflammatory measures. Levine plugs those numbers into the computer, and the algorithm does the rest. People with a biological age lower than their chronological age have a lower mortality risk, while those aging older from a biological standpoint have a higher mortality risk and are potentially more prone to developing the diseases associated with the higher age range. But perhaps what's most important here – unlike results from genetic testing – is that these are measures that can be changed. Doctors can take this information and empower patients to make changes to lifestyle, diet, exercise and sleep habits, and hopefully take steps to lower the risk and improve their biological age.

FCM Model Summary-Essentially, everyone has two ages: a chronological age, how old the calendar says you are, and a phenotypic or biological age, basically the age at which your body functions as it compares to average fitness or health levels.We basically take those and combine them using different algorithms to get what we call someone's phenotypic age or biological age." "People of the same chronological age aren't all at the same risk for developing cardiovascular disease or cancer or even dying," Levine said. "What [the biological age] does is actually give us a better idea of Levine and her team identified nine biomarkers taken in a simple blood test that seemed to be the most influential on lifespan. The biomarkers include blood sugar, kidney and liver measures, and immune and inflammatory measures. Levine plugs those numbers into the computer, and the algorithm does the rest. People with a biological age lower than their chronological age have a lower mortality risk, while those aging older from a biological standpoint have a higher mortality risk and are potentially more prone to developing the diseases associated with the higher age range.But perhaps what's most important here – unlike results from genetic testing – is that these are measures that can be changed. Doctors can take this information and empower patients to make changes to lifestyle, diet, exercise and sleep habits, and hopefully take steps to lower the risk and improve their biological age.

Human Genarated Summary-People with a biological age lower than their chronological age have a lower mortality risk, while those aging older from a biological standpoint have a higher mortality risk and are potentially more prone to developing the diseases associated with the higher age range. Essentially, everyone has two ages: a chronological age, how old the calendar says you are, and a phenotypic or biological age, basically the age at which your body functions as it compares to average fitness or health levels. People of the same chronological age aren't all at the same risk for developing cardiovascular disease or cancer or even dying," Levine said. In my lab, we work on a lot of different types of aging measures," Levine said. Doctors can take this information and empower patients to make changes to lifestyle, diet, exercise and sleep habits, and hopefully take steps to lower the risk and improve their biological age. What [the biological age] does is actually give us a better idea of Levine and her team identified nine biomarkers taken in a simple blood test that seemed to be the most influential on lifespan. The biomarkers include blood sugar, kidney and liver measures, and immune and inflammatory measures.

CHAPTER 5

CONCLUSION

Text classification and summarization has been one of the core area of NLP. NLP was developed under the intent for easier and effective computer human interaction and it has produced a lot of worthy research areas such as text summarization. Many efforts have been put into action to improve the technique of summarizing the text efficiently to get advantage over the increasing data all around. This huge data needs to be tracked and text summarization is the procedure to elevate the data tracking task. As a novel evidence of effort, a new approach to extractive text summarization technique has been presented in this paper. Our Fuzzy C-Means clustering model which is based on the combination of six most significant sentence ranking features has provided a new dimension to the extractive text summarization technique in significant way. We have taken the ranking features considering the fact that they are the most effective and useful for extractive text summarization and they have given our expected result in the study. The FCM model has proved to generate the main ideas of the document. We have implemented the technique on the CNN dataset to test the result to evident our claim of improving the technique. We also found out that the F-measure is highest for our chosen algorithm Fuzzy C-Means, compared to others. This paper has made a bold impression on overall text summarization technique by its uniqueness and strategies. In future, the ideas can be extended and applied also for sentence ranking procedure in abstractive text summarization technique and take a step closer to attain similar positive results as for extractive approach.

BIBLIOGRAPHY

[1] R. Ferreira, L. de Souza Cabral, R. D. Lins, G. P. e Silva, F. L. G. de Freitas, G. D. C. Cavalcanti, R. Lima, S. J. Simske, and L. Favaro, "Assessing sentence scoring techniques for extractive text summarization," *Expert Syst. Appl.*, vol. 40, pp. 5755–5764, 2013.

[2] E. Hovy, "Text summarization," in *The Oxford Handbook of Computational Linguistics 2nd edition*, 2003.

[3] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," in *Computer, Communication and Signal Processing (ICCCSP), 2017 International Conference on*, pp. 1–6, IEEE, 2017.

[4] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[5] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[6] N. N. . A. S. A. Rahman, A. M. Muntasir ; Saleheen, "Automatic text summarization using fuzzy c means clustering," 2017.

[7] D. Das and A. F. Martins, "A survey on automatic text summarization," *Literature Survey for the Language and Statistics II course at CMU*, vol. 4, pp. 192–195, 2007.

[8] P. B. Baxendale, "Machine-made index for technical literature,Äîan experiment," *IBM Journal of Research and Development*, vol. 2, no. 4, pp. 354–361, 1958.

[9] G. Rath, A. Resnick, and T. Savage, "The formation of abstracts by the selection of sentences. part i. sentence selection by men and machines," *American Documentation*, vol. 12, no. 2, pp. 139–141, 1961.

[10] I. Mani, *Advances in automatic text summarization*. MIT press, 1999.

[11] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 68–73, ACM, 1995.

[12] G. Salton, A. Singhal, M. Mitra, and C. Buckley, "Automatic text structuring and summarization," *Information processing & management*, vol. 33, no. 2, pp. 193–207, 1997.

[13] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of artificial intelligence research*, vol. 22, pp. 457–479, 2004.

[14] L. A. Zadeh *et al.*, "Fuzzy sets," *Information and control*, vol. 8, no. 3, pp. 338–353, 1965.

[15] L. Suanmali, N. Salim, and M. S. Binwahlan, "Feature-based sentence extraction using fuzzy inference rules," in *2009 International Conference on Signal Processing Systems*, pp. 511–515, IEEE, 2009.

[16] L. Suanmali, N. Salim, and M. S. Binwahlan, "Fuzzy logic based method for improving text summarization," *arXiv preprint arXiv:0906.4690*, 2009.

[17] R. Ferreira, F. Freitas, L. de Souza Cabral, R. D. Lins, R. Lima, G. França, S. J. Simskez, and L. Favaro, "A four dimension graph model for automatic text summarization," in *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM*

International Joint Conferences on, vol. 1, pp. 389–396, IEEE, 2013.

[18] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," 1973.

[19] H.-T. Zheng, S.-Q. Gong, J.-M. Guo, and W.-Z. Wu, "Exploiting conceptual relations of sentences for multi-document summarization," in *International Conference on Web-Age Information Management*, pp. 506–510, Springer, 2015.

[20] Y. K. Meena and D. Gopalani, "Domain independent framework for automatic text summarization," *Procedia Computer Science*, vol. 48, pp. 722–727, 2015.

[21] Y. Sankarasubramaniam, K. Ramanathan, and S. Ghosh, "Text summarization using wikipedia," *Information Processing & Management*, vol. 50, no. 3, pp. 443–461, 2014.

[22] K. V. Kumar and D. Yadav, "An improvised extractive approach to hindi text summarization," in *Information Systems Design and Intelligent Applications*, pp. 291–300, Springer, 2015.

[23] T. S. K. P. Corps, "News article writing pattern," 2011.

[24] C.-Y. Lin, "Training a selection function for extraction," in *Proceedings of the eighth international conference on Information and knowledge management*, pp. 55–62, ACM, 1999.